Evidenzbasierte Hochschuldidaktik

Ingrid Scharlau & Tobias Jenert

Zusammenfassung: *Das Kapitel wirft einen kritischen Blick auf Evidenzbasierung. Es zeigt, wie sich methodische Probleme bzw. fragwürdige methodische Praktiken auf die Aussagekraft empirischer Untersuchungen auswirken und damit die Idee der Evidenzbasierung gefährden. Es diskutiert die mit dem Konzept einhergehenden, impliziten epistemologischen, ontologischen und professionsbezogenen Vorstellungen über Wissen, Welt und Praxis, die hochschuldidaktischem Handeln möglicherweise nicht entsprechen. Schließlich zeigt es auf, dass die hochkomplexe und relevante Beziehung zwischen Daten, Evidenz und Argumenten im Konzept der Evidenzbasierung verkürzt wird. Ziel ist es, deutlich zu machen, welche Annahmen, Voraussetzungen und Grenzen mit dem vermeintlich neutralen Konzept für die Hochschuldidaktik verbunden sind.*

Schlagworte: *Evidenz, Evidenzbasierung, Daten, Argument, Begründung*

1 Einleitung

Die Diskussion darüber, was hochschuldidaktische Forschung ausmacht, ist nicht neu, wurde im letzten Jahrzehnt aber deutlich belebt. Die »Rechtfertigungssituation« (Salden, 2019, S. 554) in der Folge des Qualitätspakts Lehre dürfte ebenso einen Grund dafür darstellen wie der anhaltende Diskurs um die Selbstverortung der Hochschuldidaktik zwischen wissenschaftlicher Disziplin und lehrbezogener Servicefunktion (Reinmann, 2019). Letztlich geht es dabei immer darum, das eigene Handeln als wirkungsvoll auszuweisen – auf eine Art und Weise, die vom jeweiligen Adressatenkreis als wissenschaftlich fundiert anerkannt wird. In diesem Diskurs tauchen zunehmend die Begriffe »Evidenz« bzw. »evidenzbasiert« auf. Offenbar wird mit Evidenzbasierung eine besonders hohe Güte des wissenschaftlichen Wirknachweises verbunden. Allerdings wird die Begrifflichkeit sehr uneinheitlich verwendet. Mitunter wird synonym von »forschungs- bzw. evidenzbasierten« Zugängen gesprochen (Böhler et al., 2019, S. 8) und jede Art von em-

pirischer Forschung als Evidenzbasierung betrachtet.[1] An anderer Stelle referenziert der Begriff auf sehr spezifische Forschungsdesigns und impliziert damit ein Forschungsverständnis der empirisch-quantitativen Lehr-/Lernforschung (Reinmann, 2019). So argumentieren Metz-Göckel et al. (2012, S. 218): »Die rationale Begründung für eine empirische Forschung zur Lehrqualität und Lehrwirkung kann im Paradigma der evidenzbasierten Lehre (eBL) gesehen werden, also in der Forderung nach Lehre auf der Basis empirischer Befunde.« Diese Aussage kann nur dann als Argument gelten, wenn eine ganze Reihe von erkenntnistheoretischen Annahmen akzeptiert und darüber hinaus zahlreiche forschungsmethodische Voraussetzungen erfüllt werden. Diese werden im hochschuldidaktischen Diskurs um Evidenzbasierung kaum offengelegt und diskutiert.

Salden (2019) und auch Reinmann (2019) weisen auf die Probleme hin, die ein allzu unvorsichtiger Gebrauch des Labels Evidenzbasierung mit sich bringt. Wir teilen die Beobachtung, dass die Begriffe im hochschuldidaktischen Diskurs häufig zu undifferenziert gebraucht werden, und eine gewisse Sorge, weil speziell der Begriff der Evidenzbasierung in seinen Ursprungskontexten sehr spezifisch belegt ist. Es besteht daher die Gefahr, dass durch seine Verwendung Versprechen gemacht werden, die sich allein aufgrund forschungsmethodischer Einschränkungen nicht einhalten lassen und – noch problematischer – dass diese Versprechen gar nicht bewusst gegeben werden.

Nach einem einleitenden Blick auf die Definition und Herkunft des Konzepts (Abschnitt 2) zeigen wir, wie sich (derzeit verbreitete) methodische Probleme bzw. fragwürdige methodische Praktiken auf die Aussagekraft empirischer Untersuchungen auswirken (Abschnitt 3), und dass mit dem Konzept implizite epistemologische, ontologische und professionsbezogene Vorstellungen über Wissen, Welt und Praxis transportiert werden, die hochschuldidaktischem Handeln möglicherweise nicht entsprechen (Abschnitt 4). Abschnitt 5 illustriert die Kritik an einem Beispiel und zeigt, dass es zudem wichtig ist, das Konzept lokal auf seine Tauglichkeit zu prüfen. Abschließend argumentieren wir, dass die hochkomplexe und relevante Beziehung zwischen Daten, Evidenz und Argumenten im Konzept der Evidenzbasierung stark verkürzt wird (Abschnitt 6). Insgesamt möchten wir keine generelle Debatte für oder wider Evidenzbasierung führen, sondern deutlich machen, welche Annahmen, Voraussetzungen und Grenzen mit dem Konzept für die Hochschuldidaktik verbunden sind.

2 Evidenzbasierung

Der Begriff der Evidenzbasierung[2] ist relativ jung. Er wurde Ende des 20. Jahrhunderts in der Medizin eingeführt, in der sich Forscher dezidiert für eine Ausbildung anhand

1 Sehr deutlich wird dies im Sammelband von Szczyrba und Schaper (2018), in dem verschiedenste qualitative und quantitative Forschungsdesigns ebenso wie deskriptive, evaluierende und korrelative Fragestellungen als Formen evidenzbasierter Hochschuldidaktik zusammengestellt werden.

2 Im Englischen – wo er herkommt – gibt es ihn übrigens nur als Adjektiv, was bedeutet, dass man die jeweilige Tätigkeit, die der Evidenz bedarf, noch spezifizieren muss, z.B. »evidence-based reasoning«.

einer systematischen und hochwertigen empirischen Überprüfung medizinischer Behandlungen aussprachen (z.B. Guyatt et al., 1992). An die Stelle der Schulung des ärztlichen Urteils über den Einzelfall sollte die Kenntnis solcher Behandlungen treten, die sich in umfassenden empirischen Untersuchungen als die wirksamsten erweisen; hierfür steht die elliptische Frage *what works?*, mit der der Ansatz häufig gekennzeichnet wird. Seit der Jahrtausendwende mehren sich Forderungen nach Evidenzbasierung auch im Bildungsbereich (z.B. Bromme et al., 2014; Hattie, 2009; Prenzel, 2009; Slavin, 2002, 2008).

Zur Ermittlung der Wirksamkeit von Interventionen hat der Ansatz der Evidenzbasierung sehr klare Vorstellungen. Die verfügbaren empirischen Daten werden streng anhand der methodischen Qualität der Untersuchungen bewertet, in denen sie erhoben wurden, was zur Hierarchisierung von Evidenzformen führt: Aus Einzelfällen abgeleitetes Wissen oder durch Berufserfahrung erworbene Expertise stehen am unteren Ende der Hierarchie, gefolgt von systematischen Beobachtungsstudien. Danach folgen experimentelle Untersuchungen, unter denen die sogenannten RCTs, *randomized controlled trials*, als besonders hochwertig angesehen werden. Der Begriff *trial* steht für eine experimentelle Untersuchung, das Adjektiv *controlled* verweist darauf, dass die Interventionsgruppe mit einer Kontrollgruppe verglichen wird, und *randomized* bedeutet, dass die Teilnehmer:innen den Gruppen oder Bedingungen zufällig zugewiesen werden, so dass die Gruppen im Prinzip keine Unterschiede aufweisen sollten. Hinzu kommt häufig noch die Erfordernis der Blindheit oder Doppelblindheit (die Untersuchungsleiter:innen und Teilnehmer:innen kennen die untersuchte Hypothese nicht). Diese Eigenschaften sind wichtig für die interne Validität, d.h. für den Schluss von gemessenen Unterschieden zwischen den Gruppen auf die Wirksamkeit der Intervention, also den Nachweis einer kausalen Wirkung.[3] So formuliert eines der bekanntesten deutschen Statistiklehrbücher aus den Sozialwissenschaften: »Randomisierte Kontrollgruppenstudien gelten deswegen als ›Goldstandard‹ wissenschaftlicher Designs, weil sie einen klaren Kausalitätsnachweis liefern können« (Döring & Bortz, 2016, S. 94).

Nach dem üblichen Verständnis von Evidenzhierarchien wird eine noch höhere Stufe von Evidenz erreicht, wenn RCTs in Metaanalysen oder standardisierten Forschungssynthesen zusammengefasst werden, die die gesamte veröffentlichte empirische Evidenz berücksichtigen und die Untersuchungen entsprechend ihrer Qualität gewichten. Metaanalysen quantifizieren Wirksamkeit über die statistische Schätzung von Effektgrößen und sind damit genauer als Synthesen, die ihre Schlüsse aus inhaltlichen Überlegungen ziehen.

Verglichen mit Einzeluntersuchungen steigern Metaanalysen aufgrund der höheren Heterogenität in den Stichproben und Operationalisierungen sowie der Möglichkeit, auch die Bedingungen (Moderatoren) für das Auftreten von Effekten bzw. Interaktionen zwischen dem Effekt und Aspekten des Untersuchungsdesigns zu ermitteln, die externe Validität. Sie bieten zudem die Möglichkeit, abzuschätzen, wie stark die Datenlage dadurch verfälscht ist, dass Studien mit statistisch signifikantem Ergebnis häufiger publiziert werden als solche ohne (*publication bias*). Metaanalysen von Metaanalysen sind

3 Die Einhaltung weiterer methodischer Standards wie Objektivität und Reliabilität von Messinstrumenten sowie die Wahl einer geeigneten (z.B. repräsentativen) Stichprobe wird vorausgesetzt.

dort nützlich, wo die Anzahl einzelner Studien zu groß ist, um in einer Metaanalyse noch bewältigt werden zu können.

Zum Verfahren der Evidenzbasierung gehört auch, die Ergebnisse nachvollziehbar darzustellen und aufzubereiten. Hierfür haben sich spezialisierte Institutionen gegründet, etwa die Campbell Collaboration (campbellcollaboration.org), die Best Evidence Encyclopedia (bestevidence.org) und das What Works Clearinghouse (whatworksclearinghouse.gov; s.a. Bellmann & Müller, 2011; C. Campbell & Levin, 2009; Slavin, 2008; Wiseman, 2010).

3 Fragwürdige methodische Praktiken in empirischen Untersuchungen

Wie eine von uns bereits vor einigen Jahren konstatierte (Scharlau, 2018), bildet die methodisch begründete Forderung nach Evidenzbasierung einen merkwürdigen Kontrast zur aktuellen Diskussion um die Qualität empirischer Untersuchungen in verschiedenen Disziplinen, in denen Evidenzbasierung besonders populär ist, etwa Medizin, Psychologie und Neurowissenschaften. Schon seit vielen Jahrzehnten wird eindringlich auf substantielle Schwächen von empirischen Untersuchungen in diesen Feldern hingewiesen, und in jüngerer Zeit haben sich Hinweise auf tiefgreifende methodische Probleme verdichtet. Für die Hochschuldidaktik liegen unseres Wissens keine spezifischen Belege für solche Probleme vor; da aber so gut wie ausgeschlossen ist, dass sie von ihnen nicht betroffen ist, diskutieren wir sie hier am Beispiel einer ihrer nächsten Bezugswissenschaften, der Psychologie.

Das Konzept der Evidenzbasierung setzt voraus, dass alle Untersuchungen zu einem Thema veröffentlicht werden und so in Metaanalysen und systematische Reviews eingehen können und dass die Untersuchungen selbst die »tatsächlichen« Prozesse unverzerrt erfassen. Genau dies ist aber nicht gegeben. Zwar besteht keine Einigkeit darüber, wie gravierend die Probleme aktuell sind, aber man kann davon ausgehen, dass fragwürdige Forschungspraktiken, wie sie im Folgenden zusammengefasst werden, die Interpretierbarkeit einzelner Untersuchungen und Schlüsse aus dem Gesamtbild beeinträchtigen oder sogar unmöglich machen.

Das für die Evidenzbasierung direkteste Problem ist der *publication bias*. Diese Verzerrung besteht darin, dass methodisch gleichwertige Untersuchungen mit statistisch signifikantem Effekt eines untersuchten Faktors häufiger eingereicht und veröffentlicht werden als solche, bei denen die Wirkung sich nicht absichern lässt. Deswegen sind Untersuchungen, die eine Wirkung einer Intervention aufzeigen, in den veröffentlichten Texten, Ergebnissen und Daten überrepräsentiert (z.B. Fanelli, 2012). Erst in den letzten Jahren sind hier Gegenmaßnahmen ergriffen worden, etwa dadurch, dass Zeitschriften explizit zur Einreichung auch von Replikationen und Untersuchungen ohne signifikantes Ergebnis auffordern. Wie wirksam diese Gegenmaßnahmen sind, lässt sich noch nicht abschätzen.

Gezeigt hat sich außerdem, dass bei der Verarbeitung von Daten verschiedene fragwürdige Praktiken (*questionable research practices*, Simmons et al., 2011) angewendet werden, die dazu führen, dass die Existenz und Größe von Effekten überschätzt werden (*inflated effect sizes*, Francis et al., 2014). Diese *questionable research practices* sind meist keine

Folge betrügerischer Absicht, sondern wohl eher auf eine ritualhafte und gedankenlose Anwendung statistischer Verfahren, die als bloße Werkzeuge verstanden werden, auf Missverständnisse der Grundlagen statistischer Tests und eine Unterschätzung der gravierenden Folgen von Datenbearbeitung zurückzuführen (Gigerenzer, 2004; Gigerenzer et al., 2004). Die verbreitetsten dieser Praktiken werden als HARKing und *p*-hacking bezeichnet.

HARKing steht als Akronym für *hypothesizing after the results are known* (Kerr, 1998). Hypothesen nach Kenntnis der Ergebnisse zu entwickeln, ist legitim, nicht aber, sie an denselben Daten, an denen sie gewonnen wurden, statistisch zu testen. Nullhypothesensignifikanztests[4] (NHSTs), wie sie bislang routinemäßig in den allermeisten Untersuchungen eingesetzt werden, setzen voraus, dass die Hypothesen nicht durch die Daten informiert sind, streng zwischen explorativen und konfirmatorischen Analysen bzw. Tests unterschieden wird und alle Aspekte der Datenerhebung vorab festgelegt werden, etwa die Anzahl an Datensätzen, die abhängigen Variablen und deren exakte Messung sowie die Weiterverarbeitung von Daten. Diese Voraussetzungen scheinen ziemlich regelmäßig nicht erfüllt zu werden. Seit einiger Zeit wird deswegen für die Präregistrierung empirischer Untersuchungen geworben, bei der Hypothesen, Analyseverfahren und Tests bereits vor der Datenerhebung veröffentlicht werden. Dies wirkt HARKing entgegen, aber auch *p*-hacking.

p-hacking[5] bedeutet, Daten nach signifikanten Ergebnissen zu durchsuchen oder so zu verändern, dass die Ergebnisse signifikant werden. Zweifelsfreie Beispiele für das Fischen nach Signifikanz sind, verschiedene Tests zu rechnen, von denen nur die signifikanten berichtet werden, oder die Datenerhebung abzubrechen, sobald Signifikanz erreicht wurde (Simmons et al., 2011). Aber auch das Entfernen vermeintlicher Ausreißer aus den Daten, die Zusammenfassung von Daten, ihre Aufteilung in Untergruppen, die dann separat getestet werden, oder der Einbezug ungeplanter Kovariate verletzen die Voraussetzungen für konfirmatorische NHSTs, die nicht durch die Daten informiert sein dürfen, und produzieren artifiziell signifikante Ergebnisse.

Fiedler (2011) hat zudem darauf hingewiesen, dass die Überschätzung von Effekten speziell in der Psychologie (und man wird vermutlich ergänzen können: in den Bildungswissenschaften) auch dadurch verstärkt wird, dass Forscher:innen bei der Entwicklung von Untersuchungen ganz allgemein zahlreiche Freiheitsgrade haben, die sie intuitiv zugunsten großer Effekte ausnutzen. Dazu gehört die Auswahl von Material, Aufgaben, Randbedingungen, Variablen und ähnlichem.

Insgesamt führen diese (und weitere) Probleme dazu, dass die aus den publizierten Untersuchungen gewinnbare empirische Evidenz von geringerer Qualität ist, als im Ansatz der Evidenzbasierung vorausgesetzt wird. Die bereits erwähnte Präregistrierung von Untersuchungen, die die späteren Freiheiten der Forscher:innen einschränkt und sicherstellt, dass klar zwischen konfirmatorischen Tests und explorativen Datenanalysen unterschieden werden kann, ist eine von verschiedenen Gegenmaßnahmen, die derzeit

4 NHSTs sind nur die verbreitetste, nicht aber die einzige Form statistischen Schließens. Auf die interessante Frage potentieller Alternativen können wir hier nicht eingehen.

5 Der Kennwert *p* gibt die Wahrscheinlichkeit für das Stichprobenergebnis (oder ein extremeres Ergebnis) unter Annahme der Nullhypothese an; er ist der zentrale Kennwert von NHSTs.

ausprobiert werden; sie kann allerdings kaum verhindern, dass nichtsignifikante Ergebnisse nicht publiziert werden. Das Beispiel der Psychologie zeigt zudem, dass eine gewisse Skepsis hinsichtlich der Veränderung fragwürdiger Praktiken angebracht ist. Seit über 60 Jahren ist bekannt, dass die Stichproben in ihren Untersuchungen zu klein sind. Dies kann nicht nur zu voreiliger Nichtablehnung der Nullhypothese führen, sondern auch zum vermeintlichen Beleg kausaler Wirkungen, die gar nicht existieren (übrigens gerade dort, wo die potentiellen Effekte klein sind oder mit Kontextfaktoren interagieren). Daran hat sich trotz eindrücklicher Warnungen nichts verbessert – eher im Gegenteil (z.B. Cohen, 1962; Vankov et al., 2014).

4 Implizite Vorstellungen und blinde Flecken im Konzept und in der Praxis von Evidenzbasierung

Wie jeder Forschungsansatz sind das Verfahren und das Konzept der Evidenzbasierung nicht neutral, sondern basieren auf bestimmten, teils implizit vorausgesetzten Vorstellungen von Welt, Wissen oder professionellem Handeln. Einige kritische Elemente derselben wollen wir im Folgenden etwas genauer analysieren (s.a. Biesta, 2010). Dazu müssen wir zunächst historisch etwas zurückgehen und kommen in späteren Abschnitten dieses Kapitels auf die Hochschuldidaktik zurück.

Der Fokus evidenzbasierter Untersuchungen liegt auf der Ermittlung von Ursache-Wirkungs-Zusammenhängen (sofern man an letztere glauben mag; diese Frage steht aber außerhalb des Rahmens des vorliegenden Artikels); sie zielen auf die Formulierung möglichst allgemeiner Gesetzmäßigkeiten (in der Übertragung eines statistischen Begriffs meist »Effekt« genannt). Diese Zielrichtung rechtfertigt die Hierarchisierung von Evidenz: Die Überzeugung, dass Kausalität nur aus experimentell gewonnenen Daten abgeleitet werden kann, ist der Schlüssel für die Überlegenheit von RCTs als Methode.

Interessanterweise ist diese Überzeugung historisch neu und hat eine widersprüchliche Geschichte. Der Begriff »Experiment« hat in den Sozialwissenschaften im 19. und 20. Jahrhundert fundamentale Wandlungen durchgemacht (z.B. Danziger, 1985 1987). Nicht nur wurden Korrelationsuntersuchungen lange als experimentell bezeichnet (z.B. Danziger, 2000, S. 331), die Frage, welche Form von Untersuchungen die stärkeren Belege für kausale Wirkungen erbringt, wurde sehr unterschiedlich beantwortet. Für nichtexperimentelle Untersuchungen im modernen Sinne, neben Korrelationsuntersuchungen etwa Beobachtungen an Ereignissen in der Lebenswelt, sprach beispielsweise, dass sie komplexe und ganzheitliche Geflechte interagierender Prozesse besser abbilden können.

Die Vorstellung, dass das Experiment die mächtigste Methode oder der Königsweg zur Ermittlung kausaler Wirkungen sei, hat sich innerhalb der Sozialwissenschaften lediglich in der Psychologie klar durchgesetzt (z.B. D. Campbell, 1969), taucht aber auch hier erst in den 1970er Jahren in Lehrbüchern auf (Winston & Blais, 1996, S. 608), obwohl *randomized controlled trials* schon seit den 20er Jahren des 20. Jahrhundert verbreitet

waren. Es wirkt fast so, als würde mit der Verwendung des Begriffs Experiment ein besonderer epistemologischer Status beansprucht (Danziger & Dzinas, 1997, S. 45).[6]

Ein genauerer Blick in die Geschichte weist zudem ambivalente Erfahrungen mit evidenzbasierten Untersuchungen aus. Die scheinbare Erfolgsgeschichte von RCTs in den Sozialwissenschaften wäre ohne wissenschaftsexterne Faktoren nicht denkbar. In den USA wurden in den 1930er und 1960er Jahren verschiedene wohlfahrtsstaatliche Maßnahmen (und im 2. Weltkrieg dann auch Maßnahmen innerhalb der Armee) empirisch überprüft. Für den Staat war es wichtig, sich dabei als effizient und objektiv darzustellen – Werte, zu denen soziale Experimente im Stil von RCTs gut passten (Dehue, 2001). Die Untersuchungen selbst erwiesen sich jedoch als weniger aussagekräftig als erhofft und wurden deswegen als Hauptstrategie in allen Fällen wieder aufgegeben.

Ungeachtet dessen wurde Evidenzbasierung, nachdem sie in der Medizin der 1990er Jahre neu erfunden wurde, nach 2000 in die Bildungswissenschaften gewissermaßen reimportiert. Damit wurden Entscheidungen, die sich schon in der Mitte des 20. Jahrhunderts als problematisch erwiesen hatten, wiederholt, ohne dass aus den Erfahrungen gelernt worden wäre.

Keine Vorannahme, sondern eher eine forschungspraktische Folge evidenzbasierter Untersuchungen ist eine hohe Bereitschaft zur Komplexitätsreduktion. Untersuchungen vom Typus von RCTs sind beispielsweise aufgrund der erforderlichen Stichprobengrößen nur machbar, wenn sie sich lediglich sehr wenigen Ursachen widmen. Dies sieht man dann auch in der Literatur: Einzelne Untersuchungen thematisieren jeweils nur einen einzelnen Aspekt oder wenige Anteile des in Frage stehenden Phänomens, oft sogar reduziert auf einen einzelnen Einfluss, allem Anschein nach getrieben von der unausgesprochenen Erwartung, dass dieser sich irgendwann in ein komplexes Gesamtbild einordnen wird, das sich aus der Synthese einzelner Untersuchungen ergeben wird.

In der Praxis wird die Synthese jedoch häufig zugunsten weiterer Einzeluntersuchungen aufgeschoben. Zudem ist mehr als fraglich, ob sich evidenzbasierte Einzeluntersuchungen additiv zu einem komplexen Wirkungsgefüge kombinieren lassen. Ein didaktisches Kursdesign, das möglichst viele der von Hattie (2009) als besonders wirkungsvoll identifizierten pädagogischen Praktiken zusammenstellt, führt nicht zwangsweise (und sogar wahrscheinlich nicht) zu einem besseren Lernen als ein Kursdesign, das ein einzelnes didaktisches Prinzip zielgruppen- und kontextangemessen konsistent umsetzt.

Den wichtigsten Grund hierfür bilden die für soziale und psychische Prozesse typischen komplexen Interaktionen zwischen Kausalfaktoren.[7] Cronbach – der nicht verdächtig ist, unkritisch gegen quantitative Ansätze Partei zu nehmen – hat schon in der

6 Der Bezug sind die Naturwissenschaften, an denen sich die Psychologie von Beginn an stark orientiert hat. Übersehen wird, dass zentrale Eigenschaften naturwissenschaftlicher Experimente wie die Formulierung quantitativ höchst präziser Hypothesen und die Ableitung der Bedingungen aus formalen Modellen in der Psychologie nicht bzw. nur rudimentär erfolgen.

7 Hierbei handelt es sich nicht nur um Ursachen, sondern auch um Einflüsse, die eine Kausalbeziehung vermitteln (Mediatoren) und solche, die diese Beziehung beeinflussen (Moderatoren). Auf dieses wichtige Thema können wir hier nicht weiter eingehen.

zweiten Welle RCT-basierter Forschung deren Konzentration auf einfache Interventionseffekte anstelle von Interaktionen von Interventionen, Personen und Situationen als grundsätzlich elliptisch kritisiert (1986, S. 94), was nicht nur für Prozesse schulischen Lernens, sondern auch für das Studieren gelten sollte. Es ist immer möglich und oft nicht unwahrscheinlich, dass die kausalen Beziehungen, die man in einer Untersuchung ermittelt hat, in anderen Kontexten anders ausfallen, etwa weil sich weitere Moderatoren oder Mediatoren in die Kausalkette einschalten. Aufgrund der notwendigen Stichprobengröße sind Untersuchungen komplexer Interaktionen praktisch kaum durchführbar, und Cronbach war schon vor Jahren pessimistisch: »When we attend to interactions, we enter a hall of mirrors that extends to infinity« (Cronbach, 1975, S. 119). Sehr ähnlich haben sich aus theoretischer Perspektive Berliner (2002) und aus methodischer Meehl (1978) geäußert.

Zur Komplexitätsreduktion tragen noch mindestens zwei weitere Faktoren bei. So werden statistische Modelle verwendet, die lineare Wirkungen voraussetzen, auch wenn unwahrscheinlich ist, dass die untersuchten Phänomene lineare Zusammenhänge enthalten. Anderer Art ist die Komplexitätsreduktion, die mit der Messung der beeinflussten Prozesse und Fähigkeiten einhergeht. So werden im Fall der Hochschuldidaktik komplexe Kompetenzen durch einzelne, oft aus pragmatischen Gründen ausgewählte Operationalisierungen mit eingeschränkter Messgenauigkeit und Konstruktvalidität repräsentiert. Dies liegt als ein erster Schritt nahe, weil sich der evidenzbasierte Forschungsprozess prinzipiell an der Richtung vom Einfachen zu Komplexen orientiert; auch hier lässt sich praktisch aber beobachten, dass die (sehr aufwendige) Entwicklung besserer Messinstrumente immer weiter aufgeschoben wird, so dass die Forschung weitaus mehr als notwendig im Modus des Vorläufigen bleibt.

Generell müssen evidenzbasierte Untersuchungen annehmen, dass das Labor – RCTs finden im Idealfall in laborähnlichen Settings statt – lebensweltliche Prozesse hinreichend erfassen kann und diese entsprechend der Laborergebnisse verändert werden können. In der Logik der Evidenzbasierung betrifft dies die externe Validität von Untersuchungen. Die oben angesprochenen komplexen Interaktionen zeigen jedoch, dass Labore nicht einen Ausschnitt lebensweltlicher Prozesse untersuchen, sondern eine spezifisch konstruierte Situation. Diese schließt nicht nur viele interessante Einflüsse aus, sondern auch die Variabilität der Auswirkungen in Abhängigkeit von situativen Faktoren und beinahe immer den Umgang der betroffenen und beteiligten Personen mit den Interventionen. Das Verhältnis von Labor und Wirklichkeit ist kompliziert: Man kann davon ausgehen, dass Forschungsergebnisse nicht einfach auf »die Realität« angewendet werden. Der vermeintlichen Anwendung geht vielmehr – oft unbemerkt – eine Veränderung der Realität selbst voraus, in der letztere auf subtile Weise den Laborbedingungen angenähert wird, etwa indem Abläufe anders verstanden und praktiziert werden (z.B. Latour, 1988).

Schließlich bringt das Konzept der Evidenzbasierung mit sich, dass die untersuchten Personen nicht am Forschungsprozess beteiligt werden und im Idealfall (das Erfordernis der Blindheit) nichts von der zu prüfenden Hypothese wissen. Wir bestreiten nicht, dass man Menschen im Prinzip so beschreiben und beforschen kann, aber es scheint uns fraglich, ob diese aus methodischen Gründen postulierte Sichtweise für die Hochschuldidaktik sachlich hilfreich ist.

Das Bild der Forscher:innen ist übrigens ähnlich enggeführt. Diese stehen unbeteiligt außerhalb des erforschten Systems; Reflexivität ist für sie nicht erforderlich. Biesta identifiziert dies als »spectator view of knowledge« (2010, S. 493) und als Beispiel einer repräsentationalen Epistemologie, in der stillschweigend angenommen wird, dass Forscher:innen die Welt im Prinzip so erfassen können, wie sie ist. Er weist auch darauf hin, dass dies in merkwürdigem Kontrast dazu steht, dass die Erkenntnis durch Experimentieren, d.h. Eingriffe, zustande kommt; vielleicht ist das aber auch kein Kontrast, sondern Ausdruck dessen, dass sich die Forscher:innen nicht als Teil der beforschten Realität ansehen.

Auch auf die Sichtweise auf professionelles Handeln wirkt sich dieser Komplex von Vorannahmen aus. In der Logik evidenzbasierter Praxis tritt dieses als Intervention auf (Biesta, 2011), d.h. als eine – möglichst globale – Produktion definierter Folgen, zumindest aber als ein einfacher Eingriff in eine externe Wirklichkeit. Wir vermuten, dass die meisten Hochschuldidaktiker:innen ein anderes Verständnis ihres Handelns haben. Anstatt davon auszugehen, dass Menschen – Forscher:innen, Praktiker:innen, Studierende – in hochschuldidaktisch relevanten Situationen Wirkungen produzieren bzw. auf Kausaleinflüsse reagieren, werden sie implizit oder explizit davon ausgehen, dass diese auf komplexe Weise mit ihrer Umwelt interagieren. Dazu zählten auch die (erwarteten) Erwartungen der Interaktionspartner (z.B. Biesta, 2011; Stark, 2017). Lehren und Lernen erscheinen dann als symbolische oder symbolisch vermittelte Interaktionen; pädagogische Situationen als offen, rekursiv und durch Reaktivität gekennzeichnet; Wirkungen als über Bedeutung vermittelt, die die Subjekte (im emphatischen Sinne) ihnen zuschreiben. Für diese Sichtweise hat der Ansatz der Evidenzbasierung keine differenzierte Sprache; weder lässt sie sich in die Untersuchungspraxis einbinden noch ihr im Nachgang, gewissermaßen korrigierend, wieder hinzufügen.

5 Ein Beispiel: Hilft ein Schreibzentrum?

Bislang haben wir die Probleme evidenzbasierter Hochschuldidaktik abstrakt diskutiert. Im Folgenden spielen wir unsere wesentlichen Argumente an einem konkreten hochschuldidaktischen Beispiel durch.

Dieses Beispiel ist eine Universität, die die Einrichtung eines Schreibzentrums in Erwägung zieht. Der Vizepräsidentin für Lehre ist aus Gesprächen bekannt, dass viele Studierende Probleme mit dem Schreiben wissenschaftlicher Arbeiten haben; sie weiß auch, dass Schreibzentren hier interessante Angebote machen. Wie diese Angebote im Einzelnen beschaffen sind, weiß sie nicht. Die Studierenden im Senat und in der Lehrkommission begrüßen die Idee nachdrücklich und haben viele konkrete Vorschläge für die Aufgaben eines solchen Zentrums. Auch die Mehrheit der Lehrenden scheint die Idee sinnvoll zu finden; einige jedoch scheinen grundsätzliche Bedenken dazu zu haben, was ein Schreibzentrum in ihren besonderen Fächern überhaupt leisten kann, die sie allerdings nicht offen äußern. Vor diesem Hintergrund möchte die Universität wissen, ob sich eine solche Investition lohnt, und braucht dazu Belege über den Nutzen der Maßnahme.

Die Frage einer Institution, ob sie ein Schreibzentrum dauerhaft finanziert, ist recht typisch für die Forderung nach Evidenzbasierung im gebräuchlichen Sinne: Es wird eine Entscheidung mit langfristigen Folgen für viele Menschen getroffen. Sie hat außerdem ein empirisches Gesicht: Würde ein Schreibzentrum das Studium wesentlich verbessern? (Natürlich heißt das nicht, dass sie nicht auch und nicht einmal, dass sie nicht besser mit anderen Verfahren beantwortet oder entschieden werden könnte; diese Frage geht aber über unseren Beitrag hinaus.)

Da nicht anzunehmen ist, dass es bereits hinreichend aufschlussreiche RCTs oder gar Metaanalysen zu dieser Frage gibt, müsste eine entsprechende Untersuchung noch durchgeführt werden. Wenn man sie auf die Schreibkompetenzen der Studierenden bezieht, legt die Struktur der Fragestellung – Würde ein Schreibzentrum die Schreibkompetenzen der Studierenden verbessern können? – ein kontrolliertes randomisiertes Forschungsdesign nahe, in dem zwei Gruppen von Studierenden beobachtet werden, von denen die eine Angebote des Schreibzentrums wahrnimmt, die andere nicht.

Einige Voraussetzungen evidenzbasierter Untersuchungen lassen sich in unserem Beispiel erfüllen. So kann die Frage der Repräsentativität der Stichprobe an einer übersichtlich großen Institution mit etwas Aufwand prinzipiell gelöst werden, und die Untersuchung wäre aufgrund der zu erwartenden eher kleinen Auswirkungen der Maßnahme, die eine sehr große Stichprobe notwendig machen, zwar teuer, aber im Prinzip nicht unmöglich.

Andere Voraussetzungen sind problematischer. Zunächst muss entschieden werden, auf welches Können und Denken sich die Aktivitäten des Schreibzentrums auswirken sollen. Bislang liegen weder ein allgemein akzeptiertes Schreibkompetenzmodell auf universitärem Niveau noch valide Messinstrumente vor. Anders als bei Konstrukten, die in der Persönlichkeitspsychologie oder in vergleichenden Bildungsstudien erfasst werden, existieren für sehr viele hochschuldidaktisch interessante Phänomene keine etablierten Messinstrumente. Deren Entwicklung ist aufwendig, und so muss man sich mit ad hoc konstruierten Instrumenten behelfen. Mit diesen werden oft Stellvertretervariablen untersucht, etwa subjektiv eingeschätzte Kompetenzzuwächse und Zufriedenheit anstelle von tatsächlichem Lernen. Beides sind interessante abhängige Variablen; die mangelnde Konstruktvalidität ist für die Beantwortung von konkreten hochschuldidaktischen Fragen allerdings problematisch, und auch geringe Messgenauigkeit beeinträchtigt die Aussagekraft der Untersuchungen.

Ein methodisch relevantes Dilemma entsteht bei der Festlegung des Messzeitpunkts: Je näher er an der Intervention liegt, umso wahrscheinlicher ist es, Wirkungen beobachten zu können; zugleich ist es aber umso unwahrscheinlicher, dass sie stabil sind. Mehrfach zu messen, ist sicherlich möglich, verzögert und verteuert aber die Entscheidung.

Untersuchungen nach Jahren sind allerdings nicht nur teuer, sondern auch intern problematisch, da mit zunehmender Zeit immer weniger Teilnehmer:innen erreicht werden können und sich die Stichproben dabei verzerren. Menschen verschwinden nicht nur zufällig aus den Untersuchungsgruppen, sondern auch aus Gründen, die die Interpretierbarkeit der Ergebnisse stark einschränken. So könnte man beispielsweise annehmen, dass Teilnehmer:innen von Schreibworkshops, denen diese weder gefallen, noch subjektiv geholfen haben, weniger bereit sind, sich an späteren Befragungen zu beteiligen. In diesem Fall wären sie in der Untersuchungsgruppe zunehmend unter-

repräsentiert, was deren Ergebnisse in Richtung Wirksamkeit verzerren würde (s.a. Zhou & Fishbach, 2016). Je nachdem welche Ursachen für den differentiellen Schwund verantwortlich sind, können Ergebnisse auch in Richtung Unwirksamkeit verzerrt werden.

Zentral wären in unserem Beispiel die Einschränkungen der internen Validität. Zwar könnte man die Teilnehmer:innen randomisiert den Untersuchungsgruppen zuweisen, aber die Untersuchung wird nicht doppelblind sein. Den untersuchten Studierenden (und den am Rande betroffenen Lehrenden) wird das Versuchsdesign auffallen. Hieraus entstehen kritische Einschränkungen der internen Validität, etwa dadurch, dass die Studierenden sich Gedanken zur Untersuchung machen oder von den Kommiliton:innen aus der Experimentalgruppe informiert werden, was zusätzlich zu den »tatsächlichen« Auswirkungen der Maßnahmen des Schreibzentrums weitere Wirkwege eröffnet, die in den Ergebnissen von direkten Einflüssen nicht unterschieden werden können. Um nur ein Beispiel für Reaktivität zu nennen: Möglicherweise erwarten die Lehrenden und Studierenden der Kontrollgruppe von sich schlechtere Leistungen, da sie ja nicht in den Genuss der Unterstützung kommen. Solche Erwartungen können sich auf die tatsächlichen Leistungen auswirken.

Von großer Bedeutung ist schließlich auch, was nach der Einführung des Schreibzentrums passieren wird und per definitionem nicht in die empirische Untersuchung einbezogen werden kann. Vermutlich wird die geplante Interventionsstudie einen gewissen Neuheitseffekt einschließen, so dass zu erwarten ist, dass die späteren Auswirkungen kleiner sind als diejenigen, die in der ursprünglichen Untersuchung erfasst wurden. Andere Folgen sind systemischer Art: Vielleicht nehmen Lehrende ihre Anstrengungen bei der Unterstützung wissenschaftlichen Schreibens zurück, was die eigentlich positiven Auswirkungen des Schreibzentrums zunichte machen könnte.

Der letztgenannte Punkt hebt einen wichtigen Aspekt hochschuldidaktischer Maßnahmen hervor: Sie sind in der Regel lokal verankert und müssen nicht nur »im Prinzip«, sondern auch in einer Institution mit ihren besonderen Studiengängen, Studierenden und Lehrenden, ihrer Geschichte und antizipierten Zukunft funktionieren. Diese stehen in evidenzbasierten Untersuchungen aber überhaupt nicht im Fokus, und sie haben auch wenig Mittel dafür, diese Besonderheiten zur Sprache zu bringen.

Die zentrale Einschränkung evidenzbasierter Entscheidungen entsteht in unserem Beispielfall daraus, dass die meisten hochschuldidaktisch interessanten Phänomene Teil eines komplexen Interaktions- und Mehrebenengefüges von Prozessen sind, von dem, wie oben bereits erwähnt, nicht nur einzelne Untersuchungen, sondern auch Metaanalysen nur einen Teil erfassen und modellieren können. Aufgrund der anzunehmenden Interaktionen (erster und höherer Ordnungen) zwischen den beobachteten und den unbeobachteten Einflüssen hat man deswegen in einer einzelnen Untersuchung nicht nur einen Teil des Ganzen erfasst, sondern einen Teil unter besonderen und nicht beschriebenen Bedingungen (Cronbach, 1975); eine Generalisierung auf andere Situationen – die ja eigentlich als der besondere Vorteil von *randomized controlled trials* angesehen wird – ist damit schwer möglich.

6 Daten – Evidenz – Argument

Evidenz allein ist noch kein Argument. Dies scheint zuweilen übersehen zu werden; vielleicht deswegen, weil der Begriff Evidenz zumindest im Deutschen eine Mehrdeutigkeit hat, die auf seine zwei sprachlichen Wurzeln zurückgeht. Dies sind lat. *evidentia* als unmittelbare, anschauliche Einsichtigkeit und Gewissheit, ein für die europäische Philosophie zentrales (und umstrittenes) Konzept, und engl. *evidence* als empirischer oder datenbezogener Wirksamkeitsnachweis. Letzteres ist heute die dominante Verwendung; allerdings scheint die unmittelbare Gewissheit von *evidentia* im Diskurs zuweilen noch mitzuschwingen. Dies deutet sich u.a. darin an, dass der Bezug von Daten und Evidenz im Konzept der Evidenzbasierung deutlich genauer ausbuchstabiert wird als der Bezug von Evidenz und Argument.

6.1 Daten und Evidenz

Der Begriff von Evidenz, mit dem der Ansatz der Evidenzbasierung operiert, ist die empirische Evidenz, d.h. die vorhandenen Daten oder Informationen, aus denen abgeleitet werden kann, ob eine spezifische Aussage zutrifft oder nicht. Dass meist auf das spezifizierende Adjektiv »empirisch« verzichtet wird, stellt eine begriffliche Unklarheit dar (Stark, 2017).

Genau besehen ist empirische Evidenz allerdings mehr als Daten, da sozialwissenschaftliche Daten (oder Information oder Fakten) nicht für sich selbst sprechen. Um zu Evidenz zu werden, müssen sie hergestellt und ausgewertet werden. Auch dies baut auf Vorannahmen auf, die häufig nicht expliziert werden; im Folgenden werden die wichtigsten davon genannt und kurz ausgeführt.

Zur Auswertung von Daten werden Datenmodelle und theoretische Modelle benötigt (Bailer-Jones, 2009). Letztere verknüpfen die konkrete Situation der Datengewinnung mit empirischen und theoretischen Bedingungen und Einschränkungen; erstere formalisieren die Unsicherheit der Datenerfassung. Wenn solche Modelle nicht expliziert werden, heißt dies nicht, dass sie nicht existieren; in statistischen Verfahren beispielsweise sind sie als deren Voraussetzungen enthalten.

In RCTs besteht die Verbindung von theoretischem und Datenmodell darin, dass Daten in Bezug auf eine Hypothese ausgewertet werden; ein statistischer Test, in der Praxis meist ein Nullhypothesensignifikanztest (NHST), soll dann Auskunft über die Richtigkeit der Hypothese liefern. Wir gehen kurz auf die Aussagekraft von Signifikanztests ein, weil sie häufig missverstanden werden. Ein signifikanter NHST bedeutet, dass unter der Annahme der Nullhypothese (meist »*The intervention does not work*«) die erhobenen Daten so unwahrscheinlich sind, dass die Nullhypothese verworfen werden kann. Über Wahrscheinlichkeit von Hypothesen angesichts der Daten – einschließlich der Alternativhypothese »*The intervention works*« – erlauben NHSTs keine Schlüsse, auch wenn es gängige Praxis ist, sie so zu interpretieren (genauere Erläuterungen finden sich z.B. bei Dienes, 2008; Gigerenzer, 2004). Die den NHSTs zugrundeliegende Statistik kontrolliert lediglich die Langzeitfehlerraten der Entscheidung gegen die Nullhypothese. Unabhängig vom konkreten statistischen Test ist deren Aussagekraft gering, wenn die statistischen Hypothesen nicht formal mit den theoretischen Annahmen verbunden sind –

genau hierauf verzichten aber viele Untersuchungen, vermutlich weil diese enge Verbindung unbekannt ist und Tests als bloße Werkzeuge angesehen werden.

Eine weitere Schwachstelle der Verwandlung von Daten in Evidenz liegt in der Notwendigkeit begründet, alle Daten zu berücksichtigen. Im Prinzip steht dafür mit Metaanalysen ein statistisches und mit systematischen Reviews ein inhaltliches Verfahren zur Verfügung. In empirischen Einzelarbeiten – die logischerweise den weitaus größten Teil der Forschung ausmachen – spielt es aber nur eine geringe Rolle. Hierfür können zwei Gründe verantwortlich gemacht werden. Der erste ist das Fehlen formaler Modelle, deren enge Kopplung zwischen Daten und Theorien auch für die Verknüpfung verschiedener Untersuchungen nutzbar gemacht werden kann. Der zweite ist in der Publikationspraxis angesiedelt. In den Einleitungen empirischer Originaluntersuchungen wird der Forschungsstand oft gar nicht mehr systematisch und vollständig aufgearbeitet, sondern nur so weit, dass eine Lücke sichtbar wird, die durch die eigene Forschung besetzt werden kann (*creating a research space;* Swales, 1990); entsprechend begrenzt sind die Antworten. Obwohl diese Praktik deutlich älter zu sein scheint als die jüngste Welle der Evidenzbasierung, passt sie doch gut zu ihr – und übrigens auch zu der eingangs angesprochenen legitimatorischen Funktion von Evidenzbasierung.

Eine letzte Komplikation entsteht daraus, dass Daten im Sinne der Evidenzbasierung zwar Unterstützung für eine Vermutung liefern, dies aber nicht mit einer Schlussfolgerung aus den Daten gleichzusetzen ist, die Reiss (2015, S. 343) als »making up one's mind« beschreibt. Hierfür genügen Belege oder Unterstützung nicht, sondern es ist notwendig, beispielsweise alternative Erklärungen für denselben Zusammenhang auszuschließen. Auch das sieht man in evidenzbasierten Artikeln in den Sozialwissenschaften nur selten – unter anderem auch deswegen, weil schon bei der Anlage von Untersuchungen nicht zwei verschiedene inhaltliche Vermutungen einander gegenübergestellt werden.

6.2 Evidenz und Argument

Ein Argument ist auch ein Beitrag zu einer Auseinandersetzung. Argumentieren bedeutet, sich mit alternativen Erklärungen, Schlussfolgerungen und Lösungen auseinanderzusetzen; es bedeutet aber in der Regel auch, sich an andere Personen zu richten und deren Weltsicht oder Sinnkonstruktionen wahrzunehmen und zu berücksichtigen.

Die Idee der Evidenzbasierung scheint uns implizit dem Modell einer Debatte zu folgen, in der die Beteiligten mit dem Ziel eines klaren Entscheids für oder gegen eine bestimmte Idee argumentieren. Sie setzt dabei stillschweigend voraus, dass die Entscheidungskriterien aller Beteiligten gleich sind und sich ihre Interpretationen und Interessen nicht substantiell unterscheiden.

Eine zentrale Komplikation entsteht im hochschuldidaktischen Kontext daraus, dass Wissenschaften oder Fachkulturen sich in ihrem Verständnis von Evidenz oder, wo dieser Begriff nicht gebraucht wird, guter Begründungen unterscheiden. Die Geschichte der mit Bildung befassten Wissenschaften liefert zahlreiche Beispiele für unterschiedliche und sogar gegensätzliche Evidenzverständnisse, so etwa die Frage nach geistes- oder naturwissenschaftlichen Zugängen zu Phänomenen, die nach qualitativen oder quantitativen Methoden, oder die Auseinandersetzungen um die relative Gewichtung

von Theorie oder Erkenntnisgewinn auf der einen und Praxis oder Anwendungswissen auf der anderen Seite. Vielleicht weniger auffällig, da nicht in großen Oppositionen gebunden, aber deswegen nicht weniger wichtig, sind Ziele wie die gehaltreiche Interpretation von Phänomenen, die Dekonstruktion etablierter Sichtweisen und die kritische Reflexion vorhandener Vorstellungen und Aufklärung über diese. Solche fachspezifischen Referenzrahmen für Erkenntnisziele sind nicht lediglich merkwürdige Sitten akademischer *tribes* (Becher & Trowler, 2001), sondern ein konstitutives Element ihrer akademischen Arbeit. Sie erlauben es Lehrenden und Lernenden, sich Lehr- und Lernpraktiken und deren Kontexte im jeweiligen fachbezogenen Referenzrahmen und mit deren Eigenlogik zu erschließen. So formuliert Ellinger (2016, S. 102) am Beispiel der evidenzbasierten Pädagogik: »Das verstehende Recherchieren bestehender Theorie, die analytische Darstellung neuer Zusammenhänge gilt kaum mehr als eigenständige wissenschaftliche oder schöpferische Forschungstätigkeit, sondern stellt allenfalls das Pflichtprogramm zur Entwicklung geeigneter Hypothesen dar [… Dann] bedarf es der nachvollziehbaren Kraft des Arguments nicht mehr.« Mit der Orientierung an den Methoden anderer Wissenschaften verliere sich, so fürchtet Ellinger, was als genuin eigene Begriffe und Argumentationsmuster verstanden werden kann: »Wesentliche Fragen können nicht mehr gestellt, zentrale Probleme nicht mehr diskutiert werden« (S. 103). Ob man ihm in dieser starken Schlussfolgerung folgt, wäre zu diskutieren; dass Evidenzhierarchien mit ihrer auf eine Dimension beschränkten Bewertung von Evidenz die Gefahr der Verengung auf bestimmte Denkweisen und Argumente mit sich bringen, ist aber plausibel.

Hochschuldidaktische Entscheidungen fallen in Situationen, in denen das Modell des aushandelnden Dialogs adäquat scheint – es sind die Sichtweisen und Interessen verschiedener Gruppen zu berücksichtigen und abzustimmen, und zwar sowohl vor als auch nach der Entscheidung. Argumente sind dann solche Aussagenkomplexe, die die Beteiligten auf die Entscheidung beziehen können und die für sie sinnvoll sind.

7 Abschluss

Evidenzbasierung ist eine Möglichkeit der Begründung hochschuldidaktischer Entscheidungen und hochschuldidaktischen Handelns. Sie ist in dieser Funktion weniger neutral als es auf den ersten Blick scheinen mag; die oft impliziten Voraussetzungen, die dieser Ansatz macht, standen im Fokus dieses Artikels. Der Wert solcher Ansätze hängt allerdings auch davon ab, in welchem Kontext sie stehen. So mag Evidenzbasierung positivere Folgen haben, wenn sie in Situationen gefordert wird, in denen Handeln in Gefahr steht, von ideologischen Programmen geleitet zu werden, als in Situationen und Institutionen, in denen empirischen Befunden prinzipiell Aufmerksamkeit geschenkt wird (z.B. Kuhl et al., 2017).

Unser Beitrag soll deswegen auch keine grundsätzliche Gegenrede zum Einsatz evidenzbasierter Forschung in der Hochschuldidaktik darstellen. Tatsächlich gibt es Fragestellungen und Forschungskontexte, die Evidenzbasierung zulassen und einfordern. Allerdings möchten wir vor einer allzu sorglosen Verwendung des Begriffs der Evidenzbasierung warnen. Diese erfüllt zwar eine legitimatorische Funktion und kann kurzfris-

tig zu positiver Aufmerksamkeit seitens Drittmittelgebern oder Entscheidungsgremien führen. Gelingt es jedoch nicht, Probleme evidenzbasierter Untersuchungen zu vermeiden – und wir haben versucht, darauf aufmerksam zu machen, dass diese vielfältig und in der aktuellen Situation der Hochschuldidaktik nicht unwahrscheinlich sind –, sind mittel- oder langfristig Probleme absehbar, weil beispielsweise Maßnahmen weniger wirksam sind, als anfänglich angenommen wurde.

Darüber hinaus kann ein strenger Anspruch, hochschuldidaktische Praxis evidenzbasiert zu betreiben, dazu führen, dass andere Arten empirischer Evidenz aus dem Blick geraten. Die Bereiche hochschuldidaktischer Praxis, die sich nicht für evidenzbasierte Forschung eignen, stünden dann in der Gefahr, in den Bereich einer nichtwissenschaftlichen Praxis verschoben zu werden. Dasselbe gilt für Ansätze, die empirische Methoden einsetzen, die in den Evidenzhierarchien den niedrigeren Stufen von Evidenzbasierung zuzurechnen wären, und deren besondere Qualitäten wie etwa die enge Verbindung mit Theorie oder den aktuellen Handlungsmöglichkeiten und -bedarfen der beteiligten Akteure damit implizit abgewertet werden. Besonders groß ist die Gefahr der impliziten Abwertung unserer Ansicht nach auch für theoretische und normenentwickelnde Forschung. Für die konkrete Entscheidungssituation und die an ihr beteiligten Personen ist die Sorge berechtigt, dass die argumentative Auseinandersetzung mit hochschuldidaktischen Fragen auf das Problem des Nachweises von Wirksamkeit verkürzt wird und dadurch einerseits Aspekte wie das Aushandeln von Entscheidungskriterien oder auch nur Phänomenbeschreibungen, die hochschuldidaktisch sehr aufschlussreich sein können, vernachlässigt werden. Keine dieser Möglichkeiten erscheint uns für die Zukunft der Hochschuldidaktik erstrebenswert.

Literatur

Bailer-Jones, D.M. (2009). *Scientific models in philosophy of science*. Pittsburgh, Pa: University of Pittsburgh Press. https://doi.org/10.2307/j.ctt5vkdnq

Becher, T. & Trowler, P.R. (2001). *Academic tribes and territories: Intellectual enquiry and the cultures of disciplines* (2nd ed.). Buckingham [u.a.]: The Society for Research into Higher Education & Open Univ. Press.

Bellmann, J. & Müller, T. (2011) (Hg.). *Wissen, was wirkt: Kritik evidenzbasierter Pädagogik*. Wiesbaden: VS. https://doi.org/10.1007/978-3-531-93296-5

Berliner, D.C. (2002). Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18–20. https://doi.org/10.3102/0013189X031008018

Biesta, G.J.J. (2010). Why ›What works?‹ still won't work: From evidence-based to value-based education. *Studies in Philosophy and Education, 29*(5), 491–503. https://doi.org/10.1007/s11217-010-9191-x

Biesta, G.J.J. (2011). Evidenz, Erziehung und die Politik der Forschung. In J. Bellmann & T. Müller (Hg.), *Wissen, was wirkt: Kritik evidenzbasierter Pädagogik* (S. 269–278). Wiesbaden: VS. https://doi.org/10.1007/978-3-531-93296-5

Böhler, Y.-B., Heuchemer, S. & Szczyrba, B. (2019). Hochschuldidaktik, Hochschullehre und Hochschullernen – wissenschaftliche Perspektiven auf ihren Stellenwert in der Hochschulentwicklung. In Y.-B. Böhler, S. Heuchemer & B. Szczyrba (Hg.), *Hoch-*

schuldidaktik erforscht wissenschaftliche Perspektiven auf Lehren und Lernen (S. 7–14). Köln: TH Köln.

Bromme, R., Prenzel, M. & Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik: Eine Analyse von Anforderungen an die Darstellung, Interpretation und Rezeption empirischer Befunde. *Zeitschrift für Erziehungswissenschaft, 17* (Suppl.), 3–54. https://doi.org/10.1007/s11618-014-0514-5

Campbell, C. & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability, 21*(1), 47–65. https://doi.org/10.1007/s11092-008-9063-x

Campbell, D.T. (1969). Reforms as experiments. *American Psychologist, 24*(4), 409–429. http://dx.doi.org/10.1037/h0027982

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Cronbach, L.J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*(2), 116–127. https://doi.org/10.1037/h0076829

Cronbach, L.J. (1986). Social inquiry by and for earthlings. In D.W. Fiske & R.A. Shweder (Eds.), *Metatheory in social science: Pluralities and subjectivities* (pp. 83-107). Chicago [u.a.]: University of Chicago Press.

Danziger, K. (1985). The origins of the psychological experiment as a social institution. *American Psychologist, 40*(2), 133–140. https://doi.org/10.1037/0003-066X.40.2.133

Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Krüger, G. Gigerenzer & M.S. Morgan (Eds.), *The probabilistic revolution Vol 2: Ideas in the sciences* (pp. 35–48). Cambridge, Mass. [u.a.]: MIT Press.

Danziger, K. (2000). Making social psychology experimental: A conceptual history, 1920–1970. *Journal of the History of the Behavioral Sciences, 36*(4), 329–347. https://doi.org/10.1002/1520-6696(200023)36:4 %3C329::AID-JHBS3 %3E3.0.CO;2-5

Danziger, K. & Dzinas, K. (1997). How psychology got its variables. *Canadian Psychology, 38*(1), 43–48. https://doi.org/10.1037/0708-5591.38.1.43

Dehue, T. (2001). Establishing the experimenting society: The historical origin of social experimentation according to the randomized controlled design. *American Journal of Psychology, 114*(2), 283-302. https://doi.org/10.2307/1423518

Dienes, Z. (2008). Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. Basingstoke, Hampshire; New York, N.Y.: Palgrave-Macmillan.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-41089-5

Ellinger, S. (2016). Ökonomisierung + Inklusion = Evidenzbasierte Pädagogik? Vom Verschwinden des Pädagogischen aus der sonderpädagogischen Forschung und Lehrerbildung. In B. Ahrbeck, S. Ellinger, O. Hechler, K. Koch & G. Schad (Hg.), *Evidenzbasierte Pädagogik: Sonderpädagogische Einwände* (S. 100–128). Stuttgart: Kohlhammer.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fiedler, K. (2011). Voodoo correlations are everywhere – not only in neuroscience. *Perspectives on Psychological Science, 6*(2), 163–171. https://doi.org/10.1177/1745691611400237

Francis, G., Tanzman, J. & Matthews, W.J. (2014). Excess success for psychology articles in the journal Science. *PLoS, ONE 9*(12), e0114255. https://doi.org/10.1371/journal.pone.0114255

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G., Krauss, S. & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, Calif.: Sage. https://dx.doi.org/10.4135/9781412986311.n21

Guyatt, G., Cairns, J., Churchill, D., et al. (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA, 268*(17), 2420–2425. https://doi.org/10.1001/jama.1992.03490170092032

Hattie, J.A.C. (2009). *Visible learning: A synthesis of over 800 meta- analyses relating to achievement.* London: Routledge. https://doi.org/10.4324/9780203887332; http://dx.doi.org/10.2304/eerj.2008.7.1.124

Kerr, N.L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kuhl, J., Gebhardt, M., Bienstein, P., Käppler, C., Quinten, S., Ritterfeld, U., Tröster, H. & Wember, F.B. (2017). Implementationsforschung als Voraussetzung für eine evidenzbasierte pädagogische Praxis. *Sonderpädagogische Förderung heute, 62*, 383–393.

Latour, B. (1988). *The pasteurization of France*. Cambridge, Mass. [u.a.]: Harvard University Press.

Meehl, P.E. (1978): Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Metz-Göckel, S., Kamphans, M. & Scholkmann, A. (2012). Hochschuldidaktische Forschung zur Lehrqualität und Lernwirksamkeit. *Zeitschrift für Erziehungswissenschaft, 15*, 213–232. https://doi.org/10.1007/s11618-012-0274-z

Prenzel, M. (2009). Challenges facing the educational system. In Standing Committee for the Social Sciences (Eds.), *Vital questions: The contribution of European social science* (pp. 30–33). European Science Foundation.

Reinmann, G. (2019). Vom Eigensinn der Hochschuldidaktik. In Y.-B. Böhler, S. Heuchemer & B. Szczyrba (Hg.), *Hochschuldidaktik erforscht wissenschaftliche Perspektiven auf Lehren und Lernen* (S. 15-26). Köln: TH Köln.

Reiss, J. (2015). A pragmatist theory of evidence. *Philosophy of Science, 82*(3), 341–362. https://doi.org/10.1086/681643

Salden, P. (2019). Evidenzbasierung in der Hochschuldidaktik: Begriff – Kontext – praktische Bedeutung. *die hochschullehre, 5*, 551–560.

Scharlau, I. (2018). Sich verständigen: Überlegungen zur Frage der Evidenzbasierung. In T. Jenert, G. Reinmann & T. Schmohl (Hg.), *Hochschulbildungsforschung. Für eine offene Zukunft der Hochschuldidaktik* (S. 105–123). Wiesbaden: Springer VS.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as signif-

icant. *Psychological Science*, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

Slavin, R.E. (2008). Evidence-based reform in education: What will it take? *European Educational Research Journal*, *7*(1), 124–128.

Stark, R. (2017). Probleme evidenzbasierter bzw.-orientierter pädagogischer Praxis. *Zeitschrift für Pädagogische Psychologie*, *31*(2), 99–110. https://doi.org/10.1024/1010-0652/a000201

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Szczyrba, B. & Schaper, N. (Hg.) (2018), Forschungsformate zur evidenzbasierten Fundierung hochschuldidaktischen Handelns. Köln: Cologne Open Science. https://cos.bibl.th-koeln.de/frontdoor/index/index/docId/675

Vankov, I., Bowers, J. & Munafò, M.R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*, *67*(5), 1037–1040. https://doi.org/10.1080/17470218.2014.885986

Winston, A.S. & Blais, D.J. (1996). What counts as an experiment? A transdisciplinary analysis of textbooks, 1930–1970. *The American Journal of Psychology*, *109*(4), 599–616. https://doi.org/10.2307/1423397

Wiseman, A.W. (2010). The uses of evidence for educational policymaking: Global contexts and international trends. *Review of Research in Education*, *34*(1), 1–24. https://doi.org/10.3102/0091732X09350472

Zhou, H. & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056